Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 10, No. 3, 2023

A Tongue Segmentation Algorithm Based on Deeplabv3+ Network Model

Weifeng Bu, Mingchuan Zhang

Henan University of Science and Technology, Luoyang 471023, China

Abstract: When collecting tongue images in an open en- vironment with a mobile portable collection device, there will be problems of different shooting angles and unstable lighting. Due to the strong mobility of the portable acquisition device, the captured images will inevitably be blurred by jitter, which further increases the difficulty of segmentation. This paper applies neural network to tongue images segmentation, and proposes a tongue images segmentation method based on deep convolutional neural network. This method is a tongue images segmentation method based on the semantic segmentation framework of DeeplabV3+. First, we modify the output category of the network. Because only the tongue region is segmented, segmentation targets can be divided into two categories when performing tongue images segmentation. One is the tongue region and the other is the background region. Then we replace the backbone network of DeeplabV3+ with a lightweight network and add an attention mechanism. Finally, we use the collected tongue images in the open environment to train the network. After the network obtains the initial segmentation result, tongue images are restored according to the same type of label, so as to obtain the required tongue images only containing tongues. The experimental results show that the method has higher segmentation accuracy for tongue images in open environment, and can better meet the needs of people for tongue images segmentation.

Keywords: Convolutional neural network; Semantic seg- mentation; Deep learning; Tongue segmentation.

1. Introduction

Artificial intelligence [1] technology is considered to be the representative of the fourth industrial revolution, and has attracted wide attention from all walks of life since it was proposed. It affects all aspects of people's lives and deeply expands the impact on people. As a method to real- ize artificial intelligence, machine learning has reached the commercial application level in fingerprint recognition and face detection, but it is difficult to make new progress. The emergence of deep learning has alleviated this dilemma to a certain extent. Deep learning [2] is a technology for realizing machine learning. Due to the rapid development in recent years, it is regarded as a learning method by itself.

Traditional Chinese medicine [3] has played an indelible role in the treatment of diseases of the Chinese nation for thousands of years, which is an important part of China's intangible cultural heritage. After a long period of clinical diagnosis experience, TCM has formed a unique medical theory system. Chinese medicine believes that every part of the human body is interconnected. The symptoms of the internal organs are all reflected on the tongue coating. By observing the tongue coating, we can understand the health status of the body. The traditional tongue diagnosis [4] is made by TCM doctors according to their own knowledge level to observe the diagnosis results with the eyes. The results of the diagnosis are easily influenced by the doctor's personal level. The diagnostic process and diagnostic experience are difficult to preserve. These factors have greatly affected the development of TCM and hindered the academic exchanges of TCM. Therefore, digital tongue diagnosis is an inevitable trend of TCM tongue diagnosis. Digital tongue diagnosis makes tongue diagnosis more objective. Objectification and Stan- dardization of Tongue Diagnosis Addresses Human Fac- tors. Digital tongue diagnosis can provide more accurate treatment plans, so that the diagnosis process and results of tongue diagnosis can be effectively preserved, which is of great significance to the development of traditional Chinese medicine.

The objective research of tongue diagnosis is divided into the processing and analysis of tongue image infor- mation. More and more image processing technologies [5] are applied in the field of digital tongue diagnosis. In tongue image processing, more accurate analysis can be obtained only when a more complete tongue body is obtained. At present, domestic and foreign experts have obtained a lot of results, but there are still many challenges in the identification and segmentation of tongue images. The current research on tongue diagnosis mainly focuses on the standard light conditions of fixed equipment. there are few researches on tongue images under natural lighting conditions in open environments. In an open environment, there are many interferences from external factors, including image clarity, complex lighting, and so on. Therefore, it is more difficult to achieve tongue image segmentation in an open environment.

The difficulty in segmenting tongue images is that different people have different tongue shapes, lengths and ways of extending their tongues. The color of some tongues is similar to that of human lips, and it is difficult to distinguish them. In some cases, the tongues closely fit the lips, and the edges are so weak that it is difficult for the human eye to distinguish them. Convolutional neural networks can often be seen in medical image processing.

Input such as skin cancer classification [6], CT sign detection [7] of new coronavirus pneumonia, etc. Convolutional Neural Networks have achieved good results. Since the 1970s, image semantic segmentation has become popular and has received extensive attention from researchers [8]. With the development of artificial intelligence, after Long [9] proposed a fully convolutional neural network, the convolutional neural network was applied in semantic segmentation, which improved the semantic segmentation problem. In order to obtain accurate tongue contour, Y Cai [10] proposed the method of using an auxiliary loss function

combined with a deep convolutional neural network, using an end-to-end segmentation model, the problems of low performance of unsupervised learning methods and trivial segmentation intervals of supervised learning are solved.

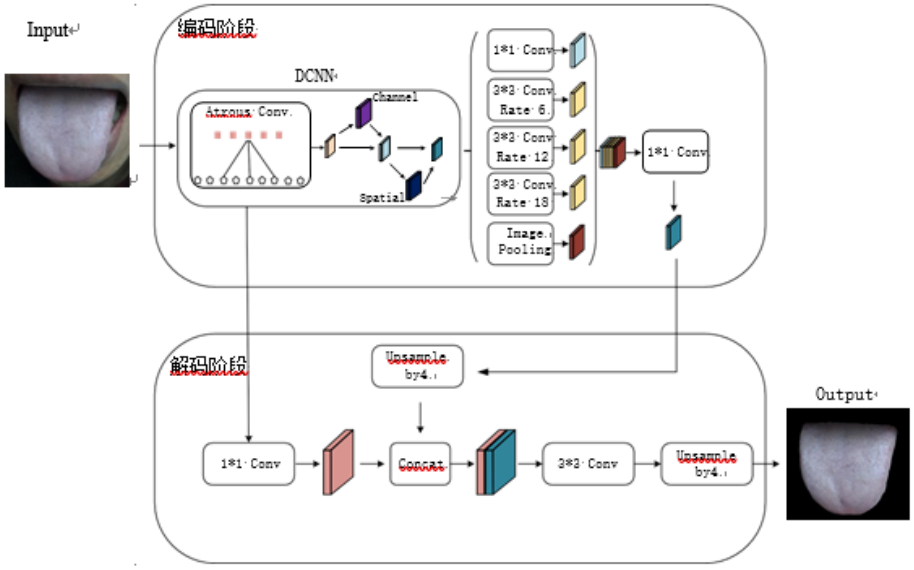

Fig. 1 The frame of tongue segmentation network

Liu [11] proposed an improved Grab Cut algorithm. Color images using the Grab Cut algorithm often have a good segmentation effect, but the disadvantage is also obvious, the disadvantage is that the segmentation time is long. The authors use a linear iterative clustering method to reduce segmentation time. Wang [12] used the HSV color space to generate the grayscale image of the image, and then used the detection operator to extract the edge information, and then used the contour generated by filtering to judge the points on the edge twice, and further determine the color range of the tongue body, and finally used H channel determines the contour of the tongue and obtains the segmentation result.

The tongue body acquired by the acquisition device in an open environment contains a large number of background areas. The quality of the segmentation effect greatly affects the analysis and judgment of the tongue image. Therefore, accurate segmentation results are an important part of digital tongue diagnosis [13]. Most of the tongue image segmentation needs to be completed with the help of human-computer interaction, which greatly reduces the automation degree of digital tongue diagnosis [14], and the segmentation effect is also unstable and the robustness is poor. Many tongue image segmentation algorithms are based on the standard light environment of fixed equipment, so it is difficult to deal with tongue images with complex lighting and complex backgrounds [15]. With the development of artificial intelligence, image processing technology is becoming more and more mature. Image processing technology using deep learning has great help and important application value for improving the ac- curacy of medical image segmentation [16]. Areas outside the body of the tongue can interfere with the analysis of the tongue image [17]. The effect of tongue segmentation seriously affects the diagnosis results of tongue image. Accurate tongue segmentation [18] is an important part of the objective research of tongue diagnosis in TCM. Most of the current tongue segmentation requires human-computer interaction to obtain a better segmentation effect, so the degree of automation is not high and the robustness is poor. Most of the tongue image segmentation is designed according to the standard light environment of fixed equipment, which is not good for tongue image processing in an open environment.

2. Algorithm model

The DeeplabV3+ [19] algorithm is a fast-semantic segmentation model. The algorithm is divided into two stages: encoding and decoding. In the coding stage, multi- scale features are introduced, and the algorithm can effectively fuse feature maps to obtain more boundary features of objects. In the decoding stage, the network restores the feature map of the encoding stage to the original image size. Then the network connects the low- level semantic features generated by the deep convolution neural network in the coding stage. The network uses 1

* 1 convolution to reduce the number of characteristic channels, so as to reduce the difficulty of training. After that, the network connects the characteristic map and the up sampling characteristic map to reduce the number of characteristic channels for further processing. The network undergoes a 3*3 convolution and an upsampling, and finally obtains a semantic segmentation model with a balance between speed and accuracy. The encoding and decoding methods of DeeplabV3+ can flexibly control the atrous convolution operation to obtain higher feature resolution.

As shown in Fig. 1, it is the network structure dia- gram of tongue image segmentation. Firstly, we set the segmentation category of the network as tongue and background. Then we add an attention mechanism to the network to make the computing resources focus on the tongue. The attention mechanism we use includes two modules, spatial and channel. This method is more accurate than the attention information obtained by using a module alone, and thus can allocate computing resources more reasonably. We apply the lightweight network to the tongue segmentation network, which reduces the parameters of the network, reduces the complexity of the tongue segmentation network, and makes the network segmentation faster. We use the manual labeling method to crop out the images containing only the tongue region, and then create a tongue segmentation dataset. We input the dataset into the tongue image segmentation network for training, and after obtaining the trained model, we start to

segment the tongue image. After the network obtains the initial segmentation result, the tongue body of the tongue image is restored according to the same type of label, so as to obtain the required tongue image containing only the tongue body.

In the coding stage of the tongue image segmentation network, in order to expand the receptive field, atrous convolution is used to extract more features of the tongue image. Atrous convolution can effectively control the resolution of tongue image feature extraction and bal- ance tongue image segmentation time and segmentation accuracy. We apply atrous convolution to tongue feature extraction to obtain more tongue feature information. formula for calculating the size of the convolution kernel of the atrous convolution is:

$$K = k(k-1)(r-1) \tag{1}$$

Where k is the original size of the convolution kernel and r is the porosity. Atrous convolution adjusts the size of the receptive field by changing the value of the atrous rate. The hole position of the hole convolution is filled with 0, and then the ordinary convolution operation is performed. Three different dilation rates of 6, 12 and 18 were used in the tongue segmentation network, respectively. Atrous convolution can expand the receptive field without losing resolution, making the semantic segmentation of tongue images more accurate. The receptive field of atrous convolution grows exponentially. The size of the receptive field obtained after atrous convolution expansion can be expressed as:

$$N_i = (2^{i+1} - 1) * (2^{i-1} - 1) \tag{2}$$

where i represents the expansion rate, i is an integer greater than or equal to 1, and when i=1, it means no ex- pansion. The convolution operation of atrous convolution applied in deep convolutional neural network is as follows:

$$N = (W - F + 2P)/S + 1 \tag{3}$$

where S is the stride of the convolution kernel. P represents the number of edge supplements. W is the width of the input tongue image. F represents the size of the convolution kernel. The deep convolutional neural network DCNN performs feature extraction first and then performs operations when processing tongue images, thereby reduc- ing the phenomenon of overfitting. The spatial and channel attention mechanism is added to the deep convolutional neural network to improve the utilization of computing resources, and its expression is:

$$\begin{cases} M_c(F) = \sigma(MLP(AP(F)) + MLP(MP(F))) \\ M_s(F) = \sigma(f([AP(F); MP(F)])) \end{cases} \tag{4}$$

The In formula (4), Mc represents the channel attention mechanism. σ is the activation operation for average pooling and max pooling. Ms is the spatial attention mechanism. The tongue image is extracted by DCNN to generate a high-level semantic feature and a low-level semantic feature. The high-level semantic features of the tongue image enter the ASPP module. ASPP enables the network to convert a tongue feature map of any size into the feature map size required for segmentation of the network model, which avoids the loss of feature map resolution caused by common pooling operations, and is conducive to improving the accuracy of tongue image segmentation. The calculation formula of the tongue feature map after pooling is as follows:

$$M = (h + 2p - f + 1) * (w + 2p - f + 1) \tag{5}$$

where h represents the height of the feature map. w is the width of the feature map. f represents the size of the convolution kernel. The activation function of the tongue image segmentation network is ReLU. This function is not only simple in form, but also improves the expression ability of the tongue segmentation model. The function has a wide acceptance threshold, and the use of unilateral suppression reduces the dependence of parameters, thereby reducing the problem of overfitting. The formula for the activation function is as follows:

$$f(x) = \sum_{i=1}^{\inf} \sigma(x - i + 0.5) \tag{6}$$

In the decoding stage of the tongue image segmentation network, the tongue image is segmented. The semantic segmentation of the tongue image is to classify the tongue image at the pixel level. The classification formula is:

$$y = \{p_i \mid p_i \in \Delta^K\}_{i=1}^{H*W} \tag{7}$$

between the predicted pixel and the label pixel. This paper uses the DeeplabV3+ segmentation framework, and on this basis, modifies the network structure and applies it to tongue image segmentation. The reconstructed network improves the segmentation accuracy and speed. The model uses multi-scale features and can effectively fuse feature maps of different sizes. The decoding part refines the feature information and improves the segmentation effect.

3. Experiment

3.1. Dataset

This paper uses the deep learning tongue image segmentation method, which avoids the steps of manually selecting features and improves the degree of automation. We first perform model training, and perform tongue image segmentation according to the trained network model. For performance evaluation, we use the classical mIoU algorithm to test the accuracy of segmentation. Since there is no public standard dataset for tongue image segmentation on the Internet, this paper constructs a tongue image dataset in an open environment according to the needs of the project. We set up the training and test

where H is the height of the tongue picture, W is the width of the tongue picture. represents the probability

simplex of K classes. The value of K in the tongue image segmentation network is 2. pi is the pixel that needs to be classified. The learning rate of the tongue image segmentation network is large in the early stage of network training, so that it can quickly reach the vicinity of the optimal point, and then the learning rate is reduced to make the model reach the local optimal value or the global optimal value. Using Nesterov optimizer to continuously adjust the learning rate, the specific expression is:

$$\text{lr} = \text{base\_lr}(1\text{step})\text{power} \tag{8}$$

Where base_lr represents the initial learning rate. lr represents the learning rate after each iteration. num is to- tal number of iterations of the tongue image segmentation network. step indicates the number of current iterations. power is the optimization parameter. From equation (4-8), it can be seen that the learning rate changes continuously with the increase of the number of iterations, making it suitable for the training of the network.

This paper uses the cross-entropy loss function to determine the training effect. This function can optimize parameters during network training. The segmentation of the network has only two classes, tongue and background, so the cross-entropy loss function is used to define the binary

classification problem. The expression is:

$$L = -[y\log(p) + (1 - y)\log(1 - p)] \quad (9)$$

In equation (4-9), P represents the probability of correct prediction of tongue image pixels. Y is the label pixel of the image. The smaller the value of L, the closer the interval sets with a ratio of 7:3. The tongue images in this dataset were taken at different times and at different locations by using different types of portable portable collection devices. The tongue images obtained are from different age groups and different genders. These tongue image datasets have different shooting angles and sizes, with different tongue shapes and tongue positions. Part of the tongue image data is shown in Fig. 2.

Fig. 2. Partial tongue images of tongue data set

The data set used in the tongue image segmentation experiment in this paper needs to be annotated at the pixel level. There are two categories of annotations, namely the tongue category and the background category. Every time we label a tongue image, we will get a label data. The size of the label data is the same as the original image size. The training model obtained in this paper has strong generalization ability. Because the datasets have different background information, different illumination angles and intensities, and different tongue shapes, the trained model can adapt to tongue images under various conditions.

3.2. Experimental environment

In the experiment, in order to improve the accuracy of the neural network, the public PASCAL VOC dataset was used to train the network to obtain primary features, and these features were used in the task of semantic segmentation of tongue images. Meanwhile, in order to speed up the training of the network, we set half of the training iterations to be frozen. A frozen state can specify that certain parameters are not updated. In this state, we train the network to only train some layers that affect the segmentation result. Second, we modify the learning rate accordingly based on the change in loss during network iterations. In the experiment, we set the segmentation class of the network as two classes, tongue body and background, and use the labeled tongue image data for training, and stop training when the loss and accuracy of the network no longer change. The experimental environment used in this paper is shown in Table 1.

Table 1. experiment settings

environment	parameters
operating system	Windows 10
frame	Caffe
language	Python
programming software	pycharm
CPU	Intel(R) Xeon(R) Silver 4210R
GPU	NVIDIA Quadro RTX 4000
RAM	64GB

In the experiment, in order to verify the effect of tongue segmentation in an open environment, we selected 4 representative tongue images. These tongue images have different tongue shapes, different thickness of tongue coating and different colors, which are suitable for testing data in an open environment. Fig. 3 shows the tongue image to be segmented. The first picture is taken at an oblique angle with less tongue area. Each picture has common problems such as red spots and tooth marks in varying degrees, which ensures the diversity of tongue images.

Fig. 3 Pictures of split test

3.3. Experimental result

The effect of tongue image segmentation needs an objective evaluation method. Subjective evaluation methods are compared through human eye observation. This method is greatly limited in practical application and has poor

stability. Therefore, the objective evaluation method that can be described quantitatively has greater value and broader development space. For the evaluation of image segmentation quality, a segmentation algorithm that can accurately evaluate the segmentation algorithm is needed in practical application, and it should have good real- time and universality. For the objective segmentation evaluation of tongue image, this paper uses mIoU as the evaluation index, which can objectively evaluate the segmentation accuracy.

Table 2. The comparison of segmentation performance

algorithm	time	mIoU
Kmeans	5.67	44.5%
GrabCut	5.33	76.3%
DeeplabV3+	0.22	90.88%
Ours	0.13	96.43%

In this paper, K-means algorithm shown in Fig. 4, grab- cut algorithm shown in Fig. 5 and DeeplabV3+ algorithm shown in Fig. 6 are used as comparative experiments. In order to better compare the segmentation results, the experiment uses pure black to fill the background, and restores the semantic segmentation results to the original tongue image color.

The segmentation speed and accuracy of each method are measured, and the measurement results are shown in TABLE II. The table shows the time and accuracy used by each method to divide an image on average. It can be seen from the table that the segmentation speed based on deep neural network method is far ahead of the segmentation algorithms based on graph and clustering. The method proposed in this chapter is also nearly twice as fast as deeplabv3 +, and the segmentation accuracy has reached the highest.

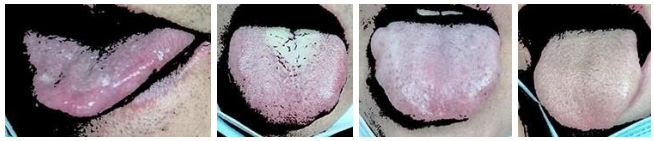

Fig. 4 K-means algorithm

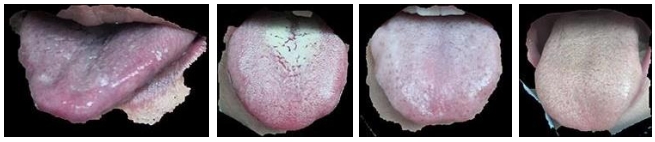

Fig. 5 Grabcut algorithm

The segmentation accuracy of each algorithm is mea- sured. The value of Miou of kmeans algorithm is 44.5%. The value of Miou of grabcut algorithm is 76.3%. The value of Miou of deeplabv3 + is 90.88%. The Miou value of the algorithm proposed in this chapter is 96.43%. From the performance comparison table of tongue image seg- mentation network, it can be seen that the segmentation algorithm proposed in this chapter is better than other methods in segmentation speed and accuracy.

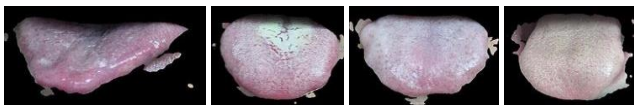

Fig. 6 DeeplabV3+ algorithm

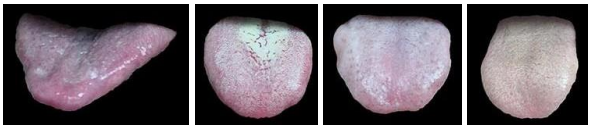

Fig. 7 Ours

4. Conclusion

The tongue segmentation algorithm model proposed in this paper is divided into two stages: encoding and decoding. In the coding stage, the hole convolution is used to extract the tongue image features, which increases the receptive field, makes up for the lost resolution information of down sampling, and improves the performance of the network. Adding attention mechanism to the network enables the network to focus computing resources on more important feature maps and feature channels, which is of great help to improve the ability of tongue feature extraction. In the decoding stage, the low-level semantic features generated by deep separable convolution are connected with the feature map generated in the coding stage, so as to obtain more detailed tongue image feature information and improve the segmentation accuracy. The tongue segmentation algorithm in this paper belongs to semantic segmentation. The feature of semantic segmen- tation is to label pixels of the same category as the same color. In this paper, the segmentation target is set as two categories: tongue body and background. The tongue body is set as white and the background is set as black. After obtaining the semantic segmentation results, we overlay the white area with the original image, so as to restore the tongue color and obtain the final tongue image image containing only the tongue.

Acknowledgments

This work was supported in part by the National Natu- ral Science Foundation of China (NSFC) under Grants No. 62002102, and in part by Henan Postdoctoral Foundation, and Foundation from the Postdoctoral Research Station of Control Science and Engineering at Henan University of Science and Technology, and in part by the Key Tech- nologies R & D Program of Henan Province under Grants No. 222102310565,232102210028 and No. 212102210088.

References

[1] Wodecki, Andrzej. Artificial intelligence methods and techniques[J]. Artificial Intelligence in Value Creation, 2019, 2: 71-132.

[2] Vitsios, Dimitrios. Prioritizing non-coding regions based on human genomic constraint and sequence context with deep learning[J]. Nature communications, 2021, 12(1): 1-14.

[3] Lin A, Chan G , Hu Y , et al. Internationalization of traditional Chinese medicine: current international market, internationalization challenges and prospective suggestions[J]. Chinese Medicine, 2018, 13(1): 1-6.

[4] Ye Y, Wei L . Design and Implementation of the Traditional Chinese Medicine Constitution System Based on the Diagnosis of Tongue and Consultation[J]. IEEE Access, 2020, (99): 4266-4278.

[5] Ku J , Harakeh A , Waslander S L . In Defense of Classical Image Processing: Fast Depth Completion on the CPU[C]. Conference on Computer and Robot Vision, 2018, 16-22.

[6] ESTEVA A, KUPREL B, NOVOA R A, et al. Dermatologist-level classification of skin cancer with deep neural networks[J]. Nature, 2017, 542(7639): 115-118.

[7] Xu T, Zhang W, Qian B. The role and challenge of artificial intelligence in new coronavirus pneumonia ct diagnosis[J]. TMR Modern Herbal Med, 2020, 3(3): 165-172.

[8] Anthimopoulos M, Christodoulidis S, Ebner L, et al. Semantic segmentation of pathological lung tissue with dilated fully convolutional networks[J]. IEEE journal of biomedical and health informatics, 2018, 23(2): 714-722.

[9] Long J, Shelhamer E, Darrell T. Fully Convolutional Networks for Semantic Segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 39(4): 640-651.

[10] Cai Y, Wang T , Liu W , et al. A robust interclass and intraclass loss function for deep learning based tongue segmentation[J]. Concurrency and Computation: Practice and Experience, 2020, 32(22): 5849.

[11] Liu B, Hu G, Zhang X, et al. Application of an improved grab cut method in tongue image segmentation[C]. International Conference on Intelligent Computing, 2018, 484-495.

[12] WANG M, ZHANG Q, ZHU J, et al. A New Computerized Tongue Diagnosis Method with Optimized Outline Extraction Algorithm Using HSV Color Model[J]. Journal of Computational & Theoretical Nanoscience, 2014, 11(6): 1556-1562.

[13] Xie J, Jing C, Zhang Z, et al. Digital tongue image analyses for health assessment[J]. Medical Review, 2021, 1(2): 172-198.

[14] Tania M H, Lwin K, Hossain M A. Advances in automated tongue diagnosis techniques[J]. Integrative Medicine Research, 2019, 8(1): 42-56.

[15] Fan S, Chen B, Zhang X, et al. Machine learning algorithms in classifying TCM tongue features in diabetes mellitus and symptoms of gastric disease[J]. European Journal of Integrative Medicine, 2021, 43: 101288.

[16] Tsai C, Lo Y, Chiang J Y, et al. Digital education and dynamic assessment of tongue diagnosis based on Mashup technique[J]. Chinese Journal of Integrative Medicine, 2017: 1-7.

[17] Zhao Q, Zhang D, Zhang B. Digital tongue image analysis in medical applications using a new tongue ColorChecker[C]. IEEE International Conference on Computer and Communications, 2016: 803-807.

[18] Fu S, Zheng H, Yang Z, et al. Computerized tongue coating nature diagnosis using convolutional neural network[C]. International Conference on Big Data Analysis, 2017: 730-734.

[19] Chen L C, Zhu Y, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]. Proceedings of the European conference on computer vision, 2018: 801-818.